Principal Component Analysis

How to Simplify and Visualise
Data Sets

Plan

- Data sets
- Curse of Dimensionality
- Struggle with Complexity
- Data sets approximation by lines and planes
- Least square definition of mean point
- “Least Square” definition of the first principal component

Plan

- Empirical covariance matrix
- Principal components are eigenvectors of empirical covariance matrix
- PCA scheme
- Eigenfaces and Eigenmuzzles

Principal components analysis (PCA) is a technique used to reduce multidimensional data sets to lower dimensions for analysis. Depending on the field of application, it is also named:

(i) the discrete Karhunen-Loève transform,
(ii) the Hotelling transform or
(iii) proper orthogonal decomposition (POD).

Everybody is a Vector

Here is a dataset

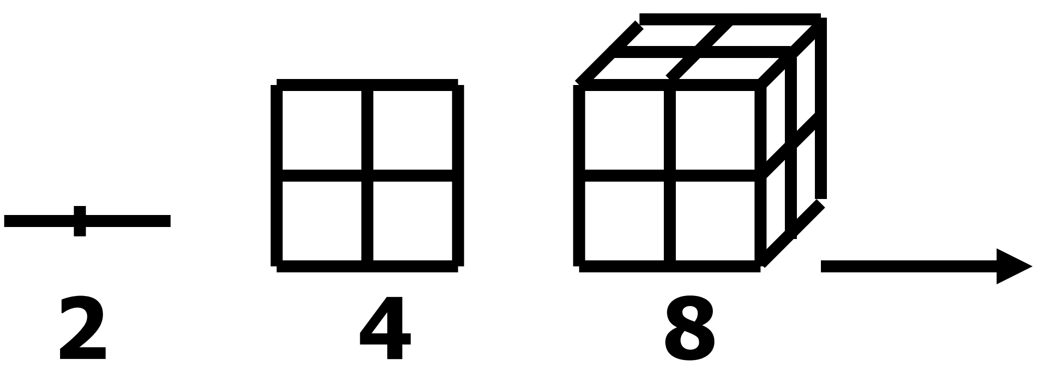
age	employme	education	edun	marital	...	job	relation	race	gender	hour	country	wealth
					...							
39	State_gov	Bachelors	13	Never_mar	...	Adm_cleri	Not_in_fan	White	Male	40	United_Sta	poor
51	Self_emp_	Bachelors	13	Married	...	Exec_man	Husband	White	Male	13	United_Sta	poor
39	Private	HS_grad	9	Divorced	...	Handlers_	Not_in_fan	White	Male	40	United_Sta	poor
54	Private	11th	7	Married	...	Handlers_	Husband	Black	Male	40	United_Sta	poor
28	Private	Bachelors	13	Married	...	Prof_speci	Wife	Black	Female	40	Cuba	poor
38	Private	Masters	14	Married	...	Exec_man	Wife	White	Female	40	United_Sta	poor
50	Private	9th	5	Married_sp	...	Other_serv	Not_in_fan	Black	Female	16	Jamaica	poor
52	Self_emp_	HS_grad	9	Married	...	Exec_man	Husband	White	Male	45	United_Sta	rich
31	Private	Masters	14	Never_mar	...	Prof_speci	Not_in_fan	White	Female	50	United_Sta	rich
42	Private	Bachelors	13	Married	...	Exec_man	Husband	White	Male	40	United_Sta	rich
37	Private	Some_coll	10	Married	...	Exec_man	Husband	Black	Male	80	United_Sta	rich
30	State_gov	Bachelors	13	Married	...	Prof_speci	Husband	Asian	Male	40	India	rich
24	Private	Bachelors	13	Never_mar	...	Adm_cleri	Own_child	White	Female	30	United_Sta	poor
33	Private	Assoc_acc	12	Never_mar	...	Sales	Not_in_fan	Black	Male	50	United_Sta	poor
41	Private	Assoc_voc	11	Married	...	Craft_repa	Husband	Asian	Male	40	*MissingV	rich
34	Private	7th_8th	4	Married	...	Transport_	Husband	Amer_India	Male	45	Mexico	poor
26	Self_emp_	HS_grad	9	Never_mar	...	Farming_fi	Own_child	White	Male	35	United_Sta	poor
33	Private	HS_grad	9	Never_mar	...	Machine_o	Unmarried	White	Male	40	United_Sta	poor
38	Private	11th	7	Married	...	Sales	Husband	White	Male	50	United_Sta	poor
44	Self_emp_	Masters	14	Divorced	...	Exec_man	Unmarried	White	Female	45	United_Sta	rich
41	Private	Doctorate	16	Married	...	Prof_speci	Husband	White	Male	60	United_Sta	rich
:	:	:	:	:	:	:	:	:	:	:	:	:

How transform them into vectors?

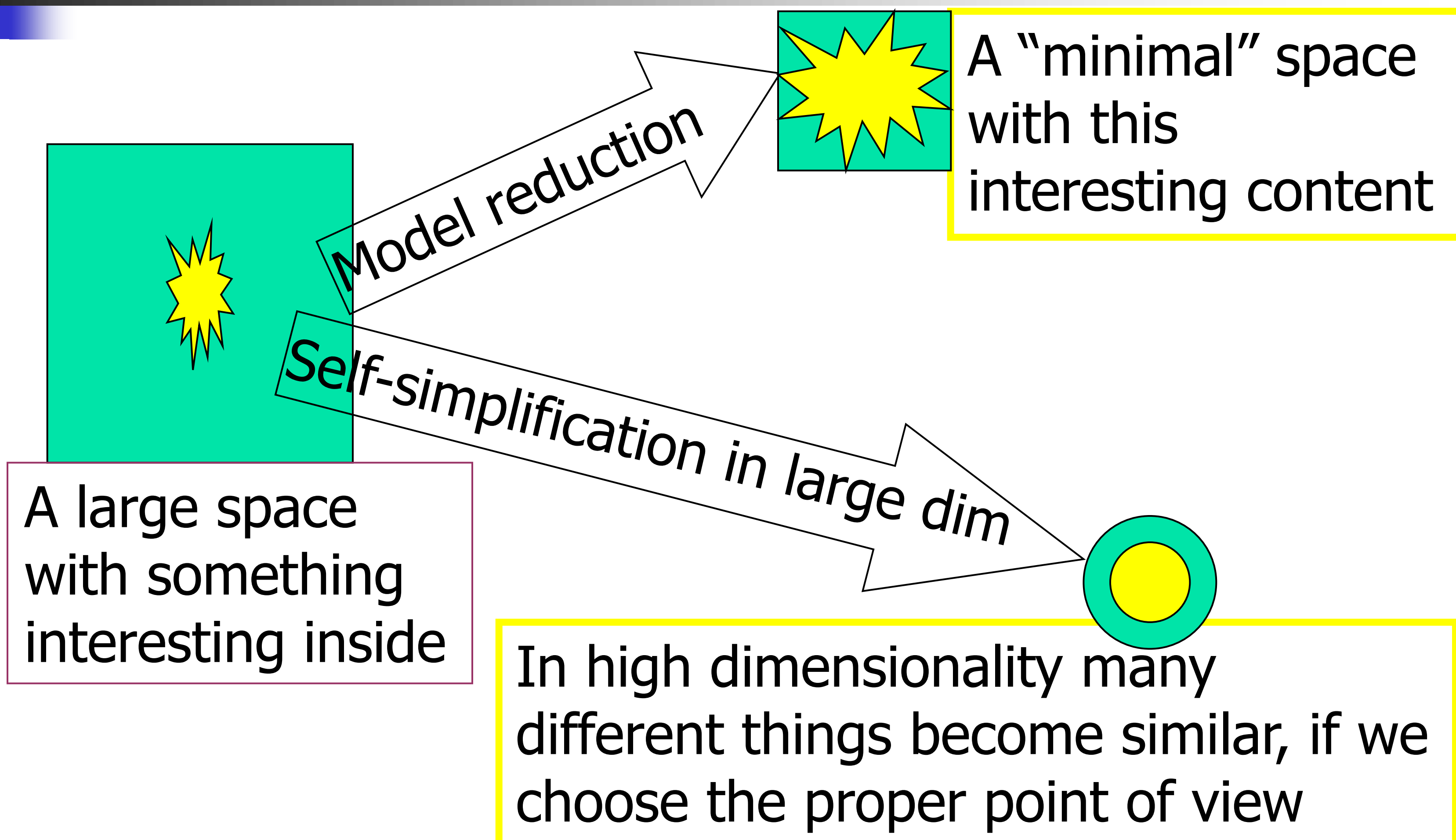
Curse of dimensionality

Curse of dimensionality (Bellman 1961) refers to the exponential growth of complexity as a function of dimensionality.

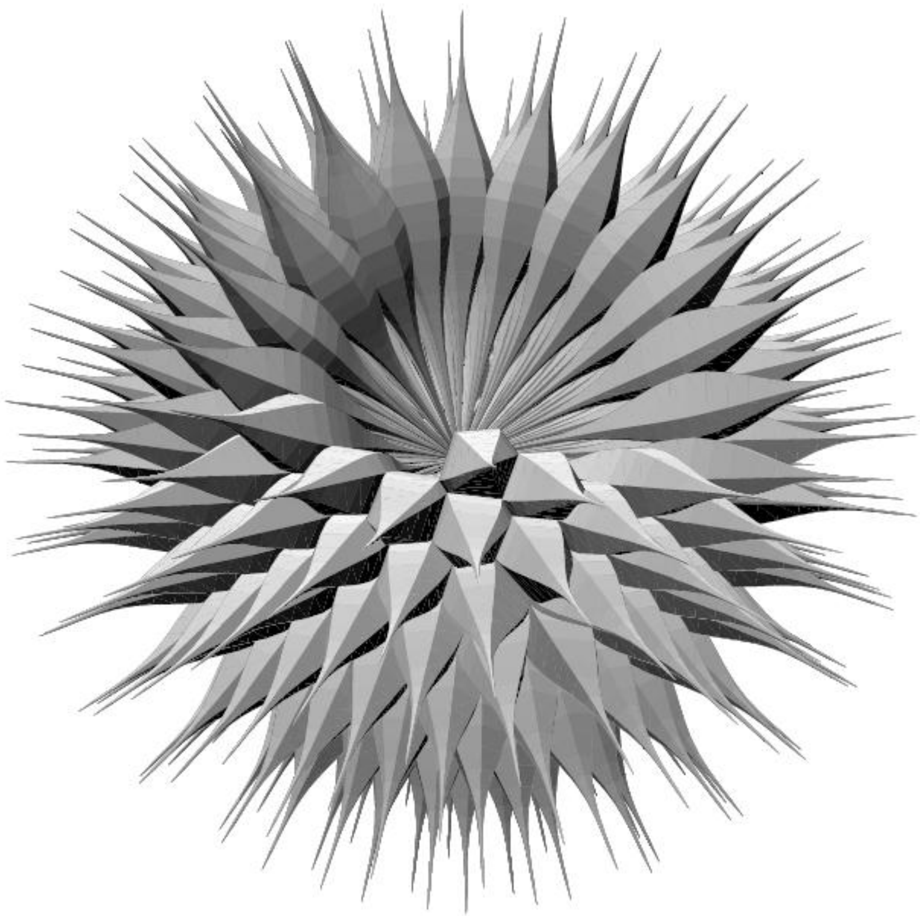
And what to do if dim>1000?



Two Main Tricks in our Struggle with Complexity



A 3D representation of an 8D hypercube



The body has the same radial distribution and the same number of vertices as the hypercube.

A very small fraction of the mass lies near a vertex.

Also, most of the interior is void.
(Hamprecht & Agrell, 2002)

Karl Pearson, 1901

LIII. *On Lines and Planes of Closest Fit to Systems of Points in Space.* By KARL PEARSON, *F.R.S., University College, London**.

(1) IN many physical, statistical, and biological investigations it is desirable to represent a system of points in plane, three, or higher dimensioned space by the "best-fitting" straight line or plane. Analytically this consists in taking

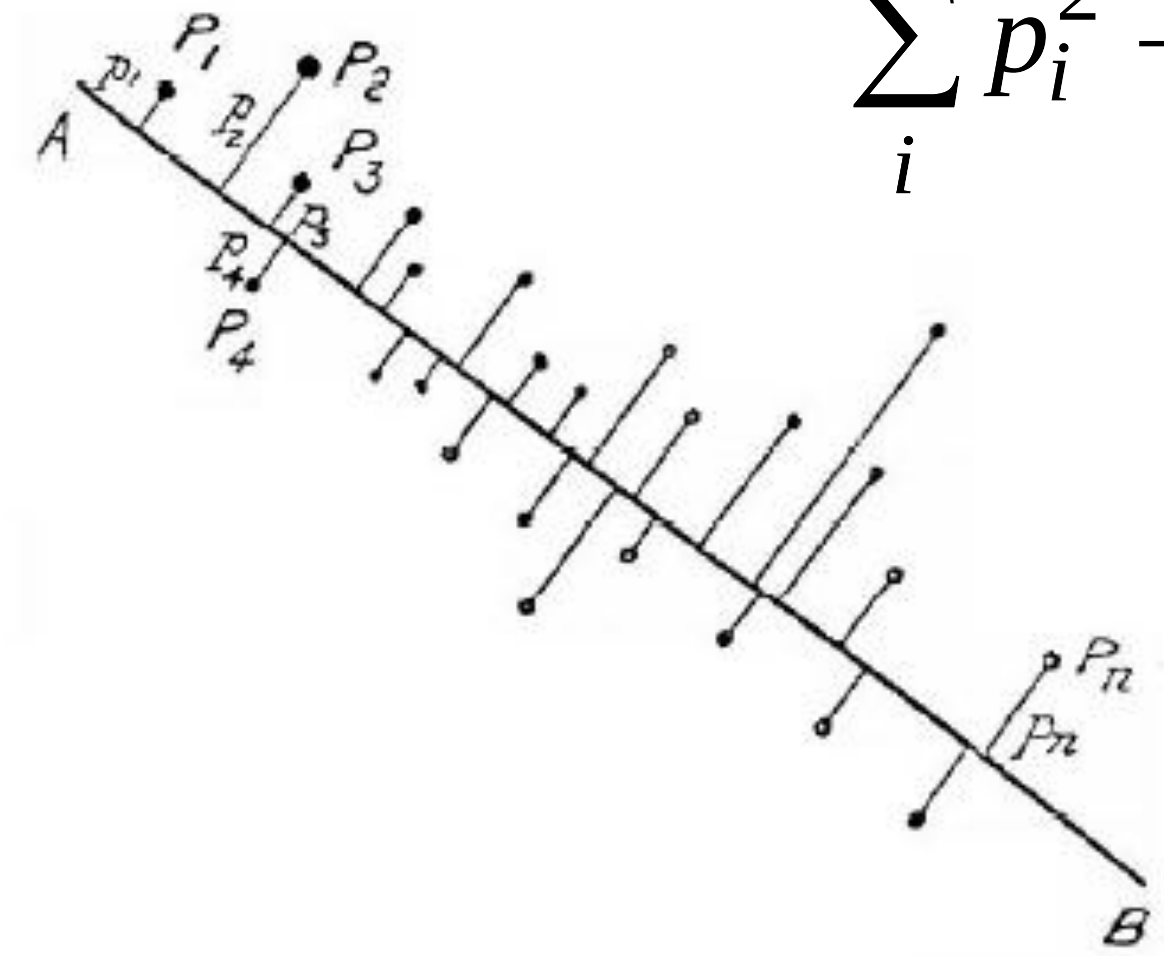
$$y = a_0 + a_1 x, \quad \text{or} \quad z = a_0 + a_1 x + b_1 y,$$

$$\text{or} \quad z = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots + a_n x_n,$$

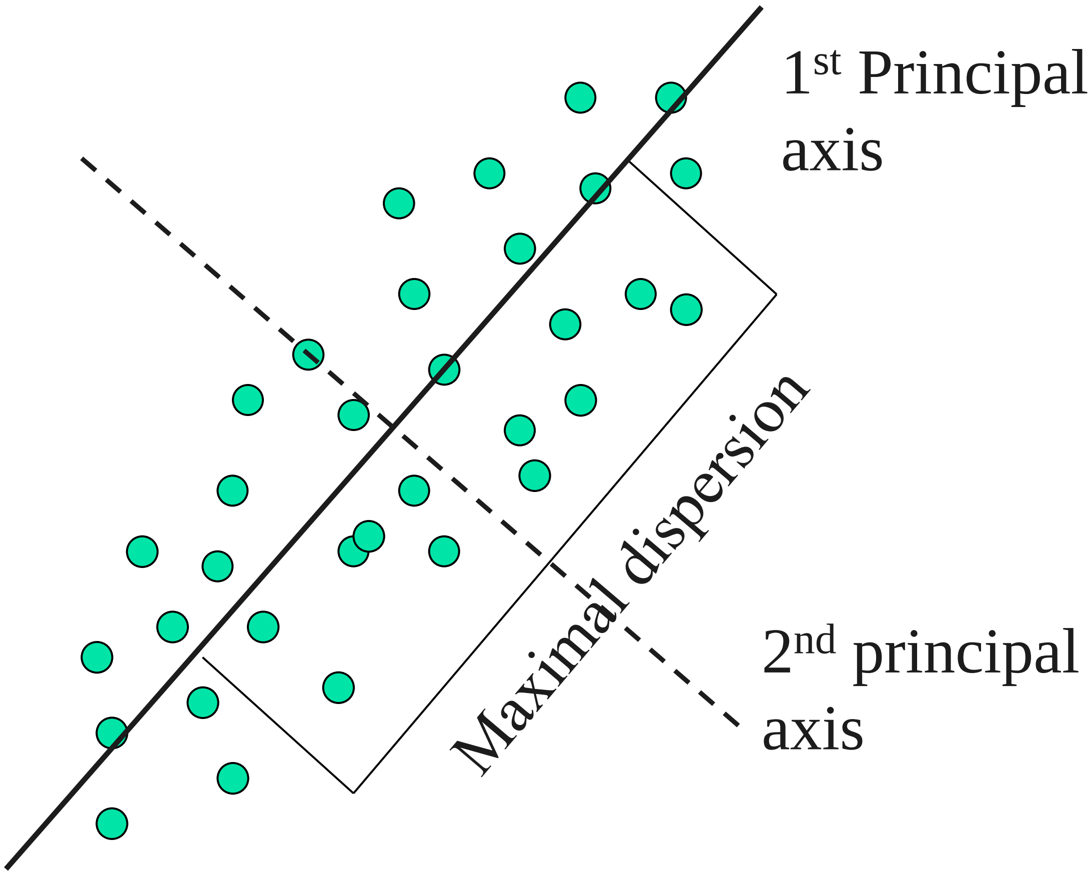
where $y, x, z, x_1, x_2, \ldots x_n$ are variables, and determining the "best" values for the constants $a_0, a_1, b_1, a_0, a_1, a_2, a_3, \ldots a_n$ in relation to the observed corresponding values of the variables. In nearly all the cases dealt with in the text-books

Data approximation by a straight line. The illustration from Pearson's paper

$$\sum_i p_i^2 \to \min$$

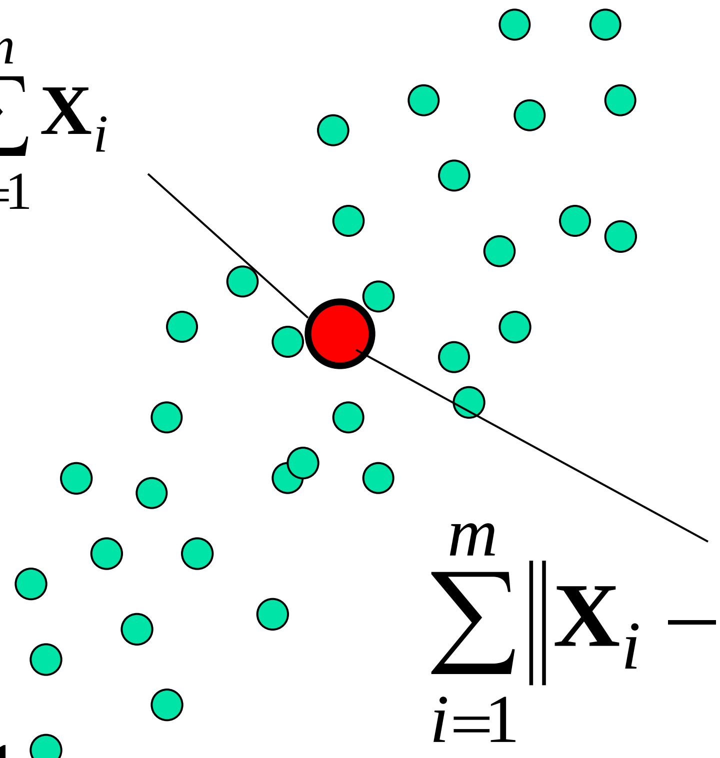


The closest approximation= The widest scattering of projections



Mean point

$$\langle \mathbf{X} \rangle = \frac{1}{m} \sum_{i=1}^{m} \mathbf{X}_i$$



$$\sum_{i=1}^{m} \left\| \mathbf{X}_i - \langle \mathbf{X} \rangle \right\|^2 \rightarrow \min$$

$\mathbf{X}_i$ – datapoints, $i = 1, \ldots m$

X_{ij} – coordinates of datapoints, $j = 1, \ldots n$

"Least Square" definition of mean point

$$\Delta^2 = \sum_{i=1}^{m} \left\| \mathbf{X}_i - \mathbf{Y} \right\|^2 \to \min, \quad \mathbf{Y} = ?$$

$$\Delta^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(X_{ij} - Y_j \right)^2 \to \min;$$

$$\frac{\partial \Delta^2}{\partial Y_j} = -2 \sum_{i=1}^{m} \left(X_{ij} - Y_j \right) = -2 \left(\left(\sum_{i=1}^{m} X_{ij} \right) - m Y_j \right) = 0;$$

$$Y_j = \frac{1}{m} \sum_{i=1}^{m} X_{ij}, \quad \mathbf{Y} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{X}_i = \langle \mathbf{X} \rangle.$$

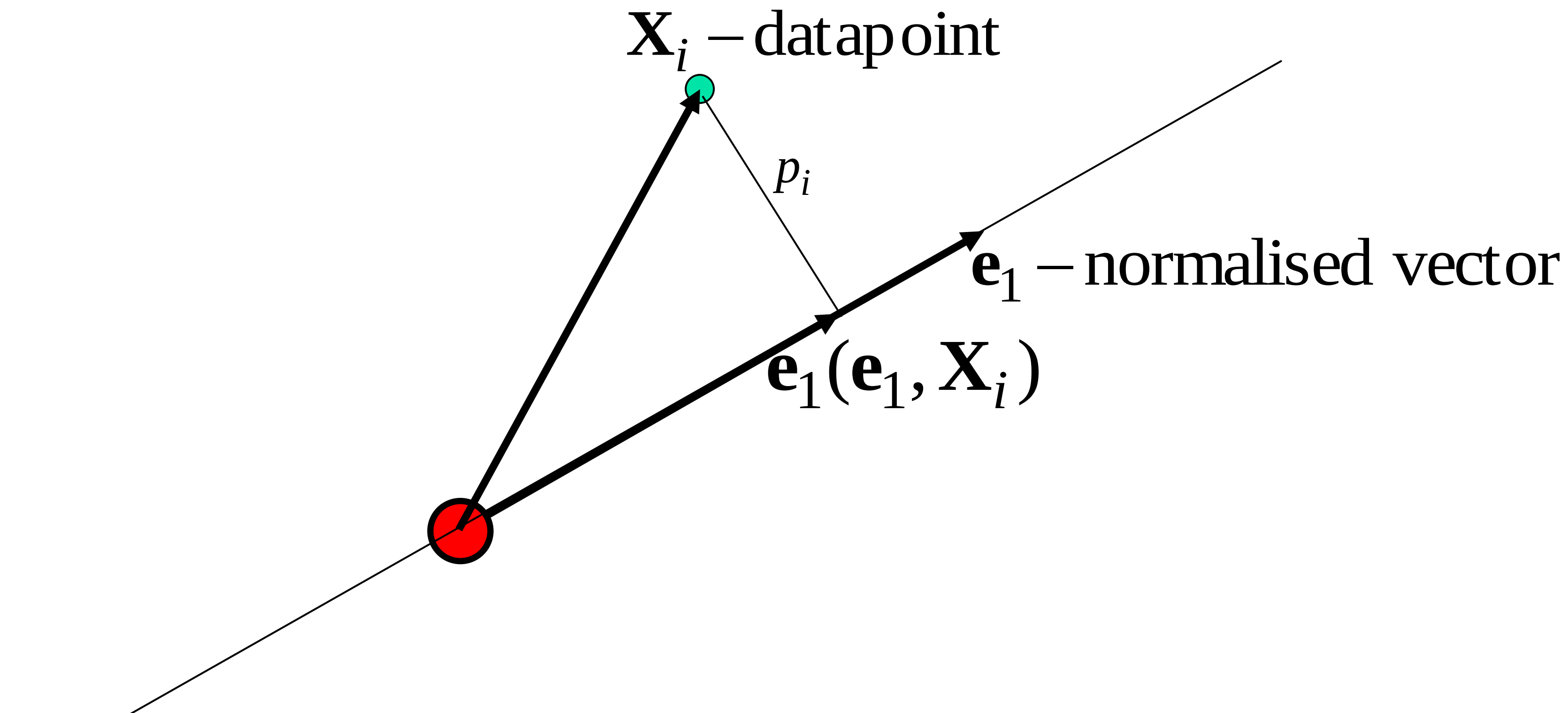
Centralisation

Let us centralise all data:

Mean Point=The Origin

$$\mathbf{X}_i \mapsto \mathbf{X}_i - \langle \mathbf{X} \rangle$$

"Least Square" definition of the first principal component



$$p_i^{\;2} = (\mathbf{X}_i - \mathbf{e}_1(\mathbf{e}_1, \mathbf{X}_i), \mathbf{X}_i - \mathbf{e}_1(\mathbf{e}_1, \mathbf{X}_i))$$

"Least Square" definition of the first principal component.2

$$\Delta^2 = \sum_{i=1}^{m} p_i{}^2 = \sum_{i=1}^{m} (\mathbf{X}_i - \mathbf{e}_1(\mathbf{e}_1, \mathbf{X}_i), \mathbf{X}_i - \mathbf{e}_1(\mathbf{e}_1, \mathbf{X}_i)) \to \min; \quad \mathbf{e}_1 = ?$$

$$\Delta^2 = \sum_{i=1}^{m} (\mathbf{X}_i - \mathbf{e}_1(\mathbf{e}_1, \mathbf{X}_i), \mathbf{X}_i - \mathbf{e}_1(\mathbf{e}_1, \mathbf{X}_i)) =$$

$$= \sum_{i=1}^{m} (\mathbf{X}_i, \mathbf{X}_i) - 2\sum_{i=1}^{m} (\mathbf{X}_i, \mathbf{e}_1)^2 + \sum_{i=1}^{m} (\mathbf{X}_i, \mathbf{e}_1)^2 = \sum_{i=1}^{m} (\mathbf{X}_i, \mathbf{X}_i) - \sum_{i=1}^{m} (\mathbf{X}_i, \mathbf{e}_1)^2;$$

$$\sum_{i=1}^{m} (\mathbf{X}_i, \mathbf{e}_1)^2 \to \max; \quad \mathbf{e}_1 = ?$$

Theorem: The closest approximation=The widest scattering of projections

"Least Square" definition of the first principal component.3

Theorem: The closest approximation=The widest scattering of projections

$$\sum_{i=1}^{m} (\mathbf{X}_i, \mathbf{e}_1)^2 \to \max; \quad \mathbf{e}_1 = ?$$

$$\sum_{i=1}^{m} (\mathbf{X}_i, \mathbf{e}_1)^2 = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} X_{ij} e_{1j} \right)^2 = \sum_{i=1}^{m} \left(\sum_{j,k=1}^{n} X_{ij} e_{1j} X_{ik} e_{1k} \right) =$$

$$= \sum_{j,k=1}^{n} e_{1j} \left(\sum_{i=1}^{m} X_{ij} X_{ik} \right) e_{1k} = m(\mathbf{e}_1, \mathbf{C}(\mathbf{X})\mathbf{e}_1),$$

where $\mathbf{C}(\mathbf{X})$ – empirical covariance matrix : $\mathbf{C}(\mathbf{X})_{jk} = \frac{1}{m} \sum_{i=1}^{m} X_{ij} X_{ik}$

Properties of empirical covariance matrix

$$\mathbf{C}(\mathbf{X})_{jk} = \frac{1}{m}\sum_{i=1}^{m} X_{ij}X_{ik}$$

1. $\mathbf{C}(\mathbf{X})$ is symmetric: $\mathbf{C}(\mathbf{X})_{jk} = \mathbf{C}(\mathbf{X})_{kj}$;

2. $\mathbf{C}(\mathbf{X})$ is positive definite: $(\mathbf{e}, \mathbf{C}(\mathbf{X})\mathbf{e}) \geq 0.$

Indeed, $(\mathbf{e}, \mathbf{C}(\mathbf{X})\mathbf{e}) = \sum_{i=1}^{m} (\mathbf{X}_i, \mathbf{e})^2 \geq 0$

Hence, eigenvalues of $\mathbf{C}(\mathbf{X})$ are non-negative real numbers,

$\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n \geq 0$

Principal components are eigenvectors of empirical covariance matrix. 1

$$\mathbf{C}(\mathbf{X})_{jk} = \frac{1}{m}\sum_{i=1}^{m} X_{ij} X_{ik}$$

Eigenvalues of $\mathbf{C}(\mathbf{X})$ are non-negative real numbers, $\lambda_1 \geq \lambda_2 \geq \dots \lambda_n \geq 0$;

$\mathbf{v}_1, \mathbf{v}_2, \dots \mathbf{v}_n$ are the correspondent orthonormal eigenvectors.

We are looking for $\mathbf{e}_1 = \sum_{i=1}^{m} \varepsilon_{1i} \mathbf{v}_i, \quad \varepsilon_{1i} = (\mathbf{e}_1, \mathbf{v}_i), \quad \sum_{i=1}^{m} \varepsilon_{1i}{}^2 = 1.$

$$\mathbf{C}(\mathbf{X})\mathbf{e}_1 = \sum_{i=1}^{m} \varepsilon_{1i} \mathbf{C}(\mathbf{X})\mathbf{v}_i = \sum_{i=1}^{m} \varepsilon_{1i} \lambda_i \mathbf{v}_i;$$

$(\mathbf{e}_1, \mathbf{C}(\mathbf{X})\mathbf{e}_1) = \sum_{i=1}^{m} \varepsilon_{1i}^2 \lambda_i \to$ max under condition $\sum_{i=1}^{m} \varepsilon_{1i}{}^2 = 1.$

Let first eigenvalues be different $\lambda_1 > \lambda_2 > \dots$

In this case, $\varepsilon_{11}^2 = 1, \varepsilon_{1i} = 0 \ (i > 1), \quad \mathbf{e}_1 = \pm \mathbf{v}_1$

Principal components are eigenvectors of empirical covariance matrix. 2

- Centralise data;
- Subtract projection on the first eigenvector;
- Solve the same minimisation problem again
- – and immediately get: $e_2=v_2$
- Iterate!

Principal components analysis

- Calculate the empirical mean
- Calculate the deviations from the mean
- Find the covariance matrix
- Find the eigenvectors and eigenvalues of the covariance matrix
- Rearrange the eigenvectors and eigenvalues
- Compute the cumulative energy content for each eigenvector
- Select a subset of the eigenvectors as low-dimensiona basis vectors
- Project the data onto the new basis